



(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) Cộng hòa xã hội chủ nghĩa Việt Nam (VN)
CỤC SỞ HỮU TRÍ TUỆ

(11)



1-0036557

(51)^{2020.01} H04M 15/00; H04M 17/00

(13) B

(21) 1-2020-03819

(22) 30/06/2020

(45) 25/08/2023 425

(43) 25/01/2021 394

(73) Tập đoàn Công nghiệp - Viễn thông Quân Đội (VN)

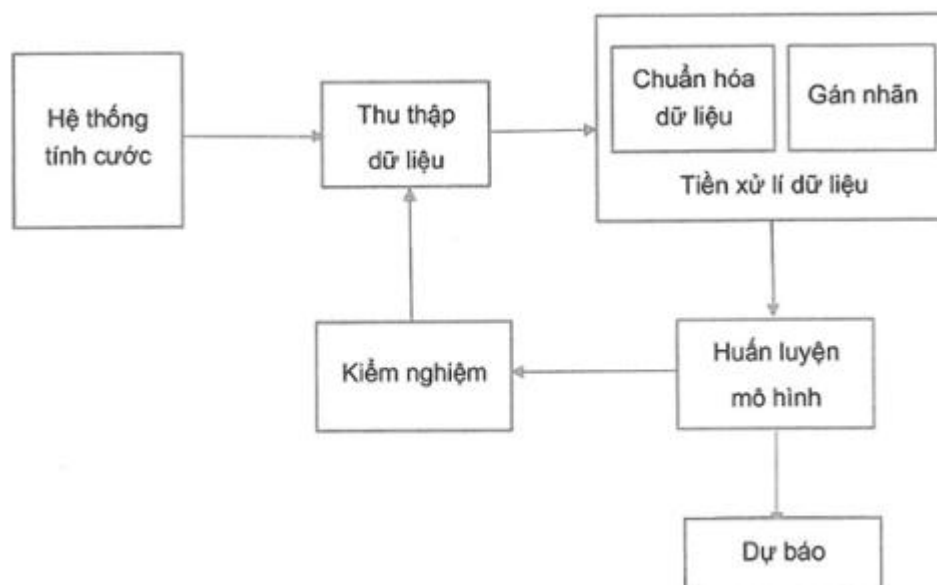
Lô D26 khu đô thị mới Cầu Giấy, phường Yên Hoà, quận Cầu Giấy, thành phố Hà Nội

(72) Phạm Công Dân (VN); Nguyễn Phi Hùng (VN); Trịnh Văn Hùng (VN); Nguyễn Đức Hải (VN).

(74) Công ty TNHH Tư vấn Quốc Dân (NACILAW)

(54) PHƯƠNG PHÁP DỰ ĐOÁN THUÊ BAO GIẢM TƯƠNG TÁC

(57) Phương pháp dự đoán thuê bao giảm tương tác giúp xử lý dữ liệu lớn trong viễn thông thu được từ hệ thống tính cước. Dữ liệu được làm sạch, chuẩn hóa theo ma trận với thuộc tính tổ chức theo cửa sổ thời gian và có tính hàng tuần, nhờ đó được dùng huấn luyện các mô hình học máy hiện đại giúp dự đoán chính xác cao thuê bao giảm tương tác, hỗ trợ đội kinh doanh tìm các biện pháp cải thiện tương tác dịch vụ viễn thông của khách hàng, giúp nâng cao doanh thu. Phương pháp bao gồm các bước: bước 1: thu thập và xây dựng cơ sở dữ liệu CDR; bước 2: tiền xử lý dữ liệu; bước 3: huấn luyện mô hình; bước 4: lựa chọn thời điểm T để kiểm nghiệm mô hình và tiến hành kiểm nghiệm mô hình; bước 5: dự đoán thuê bao giảm tương tác.



Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến phương pháp dự đoán thuê bao giảm tương tác. Phương pháp dự đoán thuê bao giảm tương tác sử dụng cửa sổ thời gian và đặc tính chuỗi thời gian của dữ liệu bản ghi chi tiết cuộc gọi (CDR) để định nghĩa, gán nhãn và xử lý, phục vụ cho việc dự đoán thuê bao giảm tương tác trong lĩnh vực viễn thông.

Tình trạng kỹ thuật của sáng chế

Hiện nay, chưa có phương pháp nào được đề xuất để xử lý khối lượng dữ liệu lớn sinh ra trong viễn thông để dự đoán thuê bao giảm tương tác. Một trong những loại dữ liệu có giá trị nhất của viễn thông là dữ liệu bản ghi chi tiết cuộc gọi sinh ra từ hệ thống tính cước, dữ liệu này trong tiếng Anh gọi tắt là CDR (Call detail record). Dữ liệu này được sinh ra liên tục khi vận hành viễn thông và do số lượng thuê bao rất lớn nên khối lượng dữ liệu tăng rất nhanh, cần có những phương pháp tốt, thiết kế hệ thống để xử lý gán nhãn và từ đó được dùng để huấn luyện các mô hình thống kê học máy.

Một số sáng chế trên thế giới về hệ thống xử lý dữ liệu viễn thông tập trung vào mục đích dự đoán thuê bao rời mạng, hoặc dự báo thuê bao giảm tương tác cho lĩnh vực năng lượng mà chưa có sáng chế xây dựng hệ thống xử lý dữ liệu và gán nhãn cho dữ liệu CDR phục vụ mục đích dự đoán thuê bao giảm tương tác trong lĩnh vực viễn thông.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là phương pháp dự đoán thuê bao giảm tương tác. Để đạt được mục đích này, phương pháp dự đoán thuê bao giảm tương tác bao gồm năm bước như sau:

Bước 1: thu thập và xây dựng cơ sở dữ liệu CDR; tại bước này, tiến hành thu thập và lưu trữ dữ liệu CDR từ hệ thống tính cước.

Bước 2: tiền xử lý dữ liệu; tại bước này, tiến hành chuẩn hóa dữ liệu và gán nhãn dữ liệu.

Bước 3: huấn luyện mô hình; chức năng của bước này là tối ưu tham số của mô hình để thu được bộ tham số tốt nhất.

Bước 4: lựa chọn thời điểm T để kiểm nghiệm mô hình và tiến hành kiểm nghiệm mô hình; trong bước này tiến hành kiểm thử độ chính xác của mô hình sau huấn luyện tại thời điểm T.

Bước 5: dự đoán thuê bao giảm tương tác; tiến hành đưa mô hình đã kiểm thử đạt vào thực tế để dự đoán thuê bao giảm tương tác.

Trong sáng chế này, phương pháp được đề xuất được thực hiện dựa trên một hệ thống bao gồm các khối như sau:

Khối thu thập và xây dựng cơ sở dữ liệu: khối này thu thập dữ liệu CDR sinh ra từ hệ thống tính cước và lưu trữ trên cơ sở dữ liệu của nền tảng xử lý dữ liệu lớn.

Khối tiền xử lý dữ liệu: khối này xử lý dữ liệu thô từ khối thu thập để chuẩn hóa và gán nhãn, lưu trữ dữ liệu được xử lý trong cơ sở dữ liệu lớn.

Khối huấn luyện mô hình: đảm nhiệm việc sử dụng dữ liệu từ khối tiền xử lý để huấn luyện mô hình học máy và lưu trữ mô hình sau khi được huấn luyện để sử dụng.

Khối kiểm nghiệm: có nhiệm vụ kiểm tra đánh giá mô hình đã huấn luyện đạt các yêu cầu chưa, kiểm tra sự cập nhật của dữ liệu, sự cần thiết phải bổ sung dữ liệu mới thu thập và huấn luyện lại để cập nhật mô hình.

Khối dự báo: khối dự báo đưa ra đánh giá mô hình đã huấn luyện đạt các tương tác, lưu các báo cáo đánh giá mô hình đã huấn luyện đạt.

Mô tả vắn tắt các hình vẽ

Hình 1 là sơ đồ khối của hệ thống xử lý dữ liệu CDR được sử dụng để ứng dụng phương pháp dự đoán thuê bao giảm tương tác;

Hình 2 là sơ đồ thể hiện chuỗi thời gian được sử dụng để dự đoán thuê bao giảm tương tác theo phương pháp được đề xuất trong sáng chế.

Mô tả chi tiết sáng chế

Phương pháp dự đoán thuê bao giảm tương tác sử dụng kỹ thuật chuỗi thời gian để chuẩn hóa dữ liệu, đưa ra phương pháp xác định khi nào một thuê bao được coi là giảm tương tác và gán nhãn cho dữ liệu. Dữ liệu được xử lý và gán nhãn sẽ được sử dụng để huấn luyện mô hình dự đoán.

Trong sáng chế này, chuỗi thời gian là dữ liệu có đặc trưng thời gian do mỗi bản ghi được thu thập ghi lại sự kiện xảy ra tại một thời điểm, chuẩn hóa dữ liệu là từ dữ liệu thô trích ra các thuộc tính cần sử dụng tổ chức lại dữ liệu dạng ma trận, giảm tương tác là đặc trưng thuê bao giảm hoạt động sử dụng dịch vụ trong một khoảng thời gian, gán nhãn được hiểu là đặt cho mỗi điểm dữ liệu gán với một giá trị 0 hoặc 1 trong đó 1 là giảm tương tác, huấn luyện mô hình dự đoán là sử dụng dữ liệu được chuẩn hóa để làm đầu vào cho mô hình, sau các vòng lặp để tinh chỉnh tham số và tối ưu mô hình học máy.

Các bước và trình tự thực hiện các bước của phương pháp này được thực hiện trên một hệ thống được trình bày trong Hình 1. Chi tiết của từng bước trong phương pháp này sẽ được mô tả cụ thể dưới đây:

- Bước 1: Thu thập và xây dựng cơ sở dữ liệu CDR;

Đầu vào của bước này là bản ghi chi tiết dữ liệu (CDR) của thuê bao người dùng. Tại bước này, cơ sở dữ liệu CDR được thu thập bằng cách viết hàm gọi dữ liệu từ hệ thống tính cước, được xây dựng bằng cách sử dụng các công nghệ sẵn có về cơ sở dữ liệu (hadoop) và tạo giao thức kết nối với hệ thống tính cước. Dữ liệu CDR sinh ra có sẵn từ hệ thống tính cước nên việc thu thập và lưu trữ trong khối thu thập và xây dựng cơ sở dữ liệu là không phức tạp.

- Bước 2: Tiền xử lý dữ liệu;

Bước này bao gồm hai bước nhỏ hơn, bao gồm:

- + B1: Chuẩn hóa dữ liệu;

Dữ liệu sau khi được thu thập và lưu trữ trực tiếp từ hệ thống tính cước thông qua khối thu thập và xây dựng cơ sở dữ liệu sẽ được tiếp tục được chuẩn hóa dữ liệu trong khối tiền xử lý dữ liệu. Chuẩn hóa dữ liệu trong sáng chế này là đưa tất cả các loại dữ liệu khác nhau của thuê bao về một mẫu chuẩn hóa duy nhất, mẫu này là một ma trận trong đó mỗi dòng thể hiện một véc-tơ thuộc tính chứa đựng các thông tin liên quan về việc sử dụng thuê bao của người dùng qua một cửa sổ thời gian n tuần. Để lập ma trận dữ liệu, sử dụng công cụ xử lý dữ liệu lớn như SparkSQL, Spark để tổng hợp dữ liệu CDR (các công cụ SparkSQL, Spark không có tên tiếng Việt sát nghĩa, vì vậy để tên gọi như hiện tại sẽ đảm bảo phạm vi bảo hộ của sáng chế hơn). CDR là bản ghi lại chi tiết thông tin sự kiện thuê bao tương tác nên từ đó ta có thể tổng hợp thông kê được cho mỗi thuê bao trong vòng 1 tuần sử dụng bao nhiêu tiền, bao nhiêu dữ liệu, tổng lượt cuộc gọi, tin nhắn trong mỗi tuần,... Mỗi trường trong dòng đó hoặc mỗi cột trong ma trận mô tả một thuộc tính hàng tuần (ví dụ tổng số tin nhắn SMSs trong vòng một tuần nào đó). Dữ liệu được chuẩn hóa có x thuộc tính trong một tuần, vì vậy một véc-tơ bao gồm thông tin sử dụng trong n tuần sẽ có $n.x$ trường. Mỗi thuộc tính tuần sẽ xuất hiện trong véc-tơ n lần, mỗi lần cho một tuần trong một cửa sổ thời gian n tuần. Dữ liệu dùng trong một cửa sổ thời gian n tuần, từ tuần a đến tuần b (với $a - b = n$), tuần $a+1$ đến tuần $b+1$, và tiếp tục, được trình bày bởi một ma trận. Các dữ liệu thuê bao sau khi được chuẩn hóa sẽ giúp phân loại các thuê bao thành hai loại: thuê bao giảm tương tác và thuê bao tương tác bình thường. Các thuê bao này được đánh nhãn hai lớp kí hiệu là 0 và 1, trong đó 1 là lớp thuê bao giảm tương tác và

0 là lớp thuê bao tương tác bình thường. Để xác định một thuê bao là giảm tương tác hay không thì từ dữ liệu chuẩn hóa, dựa trên thời lượng gọi của tuần đó thấp hơn ít nhất 40% thời lượng gọi của tuần liền trước. Để xử lý khi có khối lượng rất lớn thuê bao và cần hệ thống chạy theo thời gian thực, có thể phân cụm thuê bao theo tương tác thành một số cụm mà mỗi cụm sự tương tác của thuê bao là tương đồng. Chẳng hạn tương đồng về hành vi tiêu dùng như dùng nhiều dữ liệu 4G, nhóm dùng nhiều tin nhắn, nhóm dùng nhiều cuộc gọi, hoặc theo thu nhập trung bình (ARPU), nhóm tiêu dùng ARPU cao, trung bình, thấp...

+ B2: Gán nhãn thuê bao;

Sau khi chuẩn hóa dữ liệu của các thuê bao, sử dụng công cụ xử lý dữ liệu lớn SparkSQL, Spark để thêm một trường nhãn vào ma trận dữ liệu, trường nhãn này có giá trị 0 hoặc 1 ứng với mỗi dòng dữ liệu đặc trưng cho mỗi thuê bao, trong đó 1 là lớp thuê bao giảm tương tác và 0 là lớp thuê bao tương tác bình thường.

- Bước 3: Huấn luyện mô hình

Mô hình ở đây là một hàm toán học với một bộ tham số ban đầu được sinh ngẫu nhiên bằng máy tính. Với dữ liệu của một thuê bao cần dự đoán và kết quả dự đoán thuộc một trong hai giá trị 0 hoặc 1 ứng với thuê bao bình thường hay thuê bao giảm tương tác đã có được ở hai bước trước, tiến hành huấn luyện mô hình dự đoán thuê bao giảm tương tác. Nhiệm vụ của việc huấn luyện mô hình là quá trình tối ưu bộ tham số để thu được bộ tham số tốt nhất, giúp mô hình dự đoán chính xác nhất. Việc tối ưu mô hình cần dữ liệu được đánh nhãn ở bước 2, kết quả dự đoán của mô hình cho dữ liệu này phải ít sai lệch nhất so với nhãn thực của dữ liệu.

Dữ liệu sau khi tiền xử lý (chuẩn hóa và gán nhãn) tiếp tục được lưu trữ trên cơ sở dữ liệu của hệ thống và sử dụng trên khối huấn luyện mô hình. Các mô hình học máy có thể lựa chọn trong số các mô hình phân lớp như rừng ngẫu nhiên tăng cường độ dốc (XGboost). XGboost là một trong những phương pháp tốt nhất hiện nay trên thế giới được sử dụng rộng rãi trong lĩnh vực phân loại và dự báo từ dữ liệu, vì thế chúng tôi không mô tả chi tiết phương pháp này. Nhờ vào phương pháp xử lý dữ liệu và gán nhãn từ bước 2, giúp đưa dữ liệu về dạng chuẩn hóa để có thể dùng cho XGboost.

Theo một phương án thực hiện phương pháp được đề xuất trong sáng chế, mỗi thành phần dự đoán được huấn luyện với một ma trận gồm 4 ma trận được kết hợp theo chiều dọc biểu diễn 4 cửa sổ liên tiếp của 8 tuần (Ví dụ tuần 1-8, tuần 2-9, tuần 3-10 và tuần 4-11). Mỗi một trong 4 ma trận dữ liệu có 500 000 dòng cho 500 000 người dùng và 568 cột, do đó ma

trận dùng huấn luyện có 2 triệu dòng và 568 cột. Lí do của việc sử dụng 4 cửa sổ trong huấn luyện là bởi tính theo mùa trong hành vi của người dùng. Việc huấn luyện cần đưa đặc tính này vào để kết quả tốt và thực tế hơn. Ngoài ra sau khi huấn luyện cho mỗi một thành phần dự đoán(predictor), ta kết hợp 5, 10 hoặc 15 thành phần dự đoán để tạo thành một mô hình tổng hợp dự đoán tốt hơn.

Việc huấn luyện mô hình được triển khai trên các máy chủ có khả năng xử lý cao khi dữ liệu thuê bao lên đến hàng triệu thuê bao. Bước này giúp tinh chỉnh tham số của mô hình và đầu ra là mô hình tối ưu dựa trên tri thức thu được từ dữ liệu chuẩn hóa, mô hình được lưu lại và được kiểm nghiệm trước khi đưa vào dự đoán thuê bao giảm tương tác.

- Bước 4: Lựa chọn thời điểm T để kiểm nghiệm mô hình và tiến hành kiểm nghiệm mô hình

Để kiểm nghiệm mô hình trên khối kiểm nghiệm, với đặc thù dữ liệu chuỗi thời gian, ta chia dữ liệu dựa trên mốc thời gian tại thời điểm T, xem hình 2. Thời điểm T được lựa chọn dựa trên tập dữ liệu được thu thập, đảm bảo 75% dữ liệu trước thời điểm T làm dữ liệu huấn luyện và 25% dữ liệu sau thời điểm T làm dữ liệu kiểm nghiệm. Hiệu quả sau huấn luyện với cách xử lý dữ liệu như mô tả cho mô hình dự đoán rừng ngẫu nhiên tăng cường, đạt tỉ số F1 xấp xỉ 80%. Theo thời gian có thể mô hình sẽ lạc hậu, vì thế hiệu năng sau khi kiểm nghiệm sẽ giảm. Lúc này dữ liệu được thu thập thêm và mô hình sẽ được cập nhật và huấn luyện lại, thời điểm T cũng sẽ được cập nhật theo qui tắc 75% - 25%. Phần 25% dữ liệu dùng để kiểm nghiệm được đưa vào mô hình và dự đoán nhãn. So sánh nhãn dự đoán bởi mô hình với nhãn thật để tính được độ chính xác của mô hình. Độ chính xác cao hơn một ngưỡng cho trước, ví dụ 80% thì mô hình đạt yêu cầu.

Sau khi lựa chọn được thời điểm T để kiểm nghiệm mô hình, tiến hành kiểm nghiệm mô hình trên hệ thống máy tính dựa vào kết quả dự đoán của mô hình cho dữ liệu kiểm nghiệm. Do sự mất cân bằng dữ liệu, chẳng hạn cho tuần 2-9, 28,7% dữ liệu có nhãn giảm tương tác. Nếu một mô hình rất xấu dự đoán toàn bộ người dùng là không giảm tương tác, khi đó mô hình vẫn đạt chuẩn xác 71,3%. Hơn nữa với mục đích giữ thuê bao người dùng, cần dự báo chính xác người dùng giảm tương tác hơn là dự báo người dùng không giảm tương tác. Việc nhầm lẫn một thuê bao tương tác thành giảm tương tác có thể dẫn tới sự kém hiệu quả của các chương trình tương tác sau này, tuy nhiên bỏ qua một thuê bao giảm tương tác có thể dẫn đến mất đi một người dùng, trực tiếp ảnh hưởng hiệu suất. Vì vậy ngoài độ chuẩn xác kiểm nghiệm mô hình sẽ dùng các số đo:

$$\text{độ chính xác} = \frac{\text{Tổng người dùng giảm tương tác được dự đoán đúng}}{\text{Tổng người dùng được dự đoán là giảm tương tác}}$$

$$\text{thu hồi} = \frac{\text{Tổng người dùng giảm tương tác được dự đoán đúng}}{\text{Tổng người dùng thực tế là giảm tương tác}}$$

Và trung bình điều hòa:

$$F1 = 2 \cdot \frac{\text{độ chính xác} \cdot \text{thu hồi}}{\text{độ chính xác} + \text{thu hồi}}$$

Mô hình đạt yêu cầu khi F1 từ 75% trở lên. Tùy vào yêu cầu thực tế có thể thay đổi ngưỡng này.

- Bước 5: Dự đoán thuê bao giảm tương tác;

Trong bước này, mô hình thu được và được lưu trữ trong bước 3 đã được kiểm nghiệm trong bước 4 được sử dụng tại khối dự đoán. Lúc này với dữ liệu thuê bao mới được đưa vào hệ thống ở thời điểm hiện tại, cần dự đoán tuần tiếp theo thuê bao có giảm tương tác hay không. Nếu trong bước 3, dữ liệu lịch sử được thu thập để huấn luyện và thu được mô hình, thì bước 5 dùng mô hình từ bước 3 để dự đoán cho dữ liệu mới chưa có trong tập dữ liệu huấn luyện. Nếu bước 3 huấn luyện mô hình đòi hỏi hệ thống máy tính mạnh để thực hiện trong một thời gian khá lâu có thể mất vài giờ thì trong bước 5 với hệ thống máy tính tương tự, việc dự đoán sẽ diễn ra nhanh chóng chỉ trong vòng khoảng 15 phút.

Hệ thống máy tính sử dụng ở bước 3 và bước 5 gồm 100 máy chủ kết nối với hệ thống xử lý dữ liệu lớn Hadoop, mỗi máy chủ có cấu hình: Hệ điều hành Redhat 6.8, RAM; 256 GB, HDD 1.5TB, SSD: 1.5TB, CPU: 48 cores, Chip: Intel(R) Xeon CPU ES-2690 v2 3.00 Ghz. Chạy trên hệ thống là công cụ Spark trong hệ thống phân tán với tham số 18Gb bộ nhớ thực thi, 10 Gb bộ nhớ quản lý tiến trình, 5 nhân thực thi và 20 máy thực hiện.

Sáng chế được mô tả chi tiết bằng cách sử dụng các phương án được mô tả ở trên. Tuy nhiên, rõ ràng là đối với người hiểu biết trung bình trong lĩnh vực sáng chế không bị giới hạn ở phương án được mô tả trong phần mô tả sáng chế. Sáng chế có thể được thực hiện ở chế độ cải biến hoặc thay đổi mà không nằm ngoài phạm vi sáng chế được xác định bởi các điểm yêu cầu bảo hộ. Vì vậy những gì được mô tả trong phần mô tả sáng chế chỉ nhằm mục đích minh họa, và sẽ không áp đặt bất kỳ giới hạn nào đối với sáng chế.

Yêu cầu bảo hộ

1. Phương pháp dự đoán thuê bao giảm tương tác gồm 5 bước:

bước 1: thu thập và xây dựng cơ sở dữ liệu CDR;

đầu vào của bước này là bản ghi chi tiết dữ liệu (CDR) của thuê bao người dùng; tại bước này, cơ sở dữ liệu CDR được thu thập bằng cách viết hàm gọi dữ liệu từ hệ thống tính cước, được xây dựng bằng cách sử dụng các công nghệ sẵn có về cơ sở dữ liệu như dữ liệu lớn và tạo giao thức kết nối với hệ thống tính cước; dữ liệu CDR sinh ra có sẵn từ hệ thống tính cước nên việc thu thập và lưu trữ trong khối thu thập và xây dựng cơ sở dữ liệu là không phức tạp;

bước 2: tiền xử lý dữ liệu;

bước này bao gồm hai bước nhỏ hơn, bao gồm:

b1: chuẩn hóa dữ liệu;

mẫu chuẩn hóa là một ma trận trong đó mỗi dòng thể hiện một véc-tơ thuộc tính chứa đựng các thông tin liên quan về việc sử dụng thuê bao của người dùng qua một cửa sổ thời gian n tuần; để lập ma trận dữ liệu, sử dụng công cụ xử lý dữ liệu lớn như SparkSQL, Spark để tổng hợp dữ liệu CDR, CDR là bản ghi lại chi tiết thông tin sự kiện thuê bao tương tác nên từ đó ta có thể tổng hợp thông kê được cho mỗi thuê bao trong vòng 1 tuần sử dụng bao nhiêu tiền, bao nhiêu dữ liệu, tổng lược cuộc gọi, tin nhắn trong mỗi tuần,... mỗi trường trong dòng đó hoặc mỗi cột trong ma trận mô tả một thuộc tính hàng tuần (ví dụ tổng số tin nhắn SMSs trong vòng một tuần nào đó); dữ liệu được chuẩn hóa có x thuộc tính trong một tuần, vì vậy một véc-tơ bao gồm thông tin sử dụng trong n tuần sẽ có $n.x$ trường; mỗi thuộc tính tuần sẽ xuất hiện trong véc-tơ n lần, mỗi lần cho một tuần trong một cửa sổ thời gian n tuần; dữ liệu dùng trong một cửa sổ thời gian n tuần, từ tuần a đến tuần b (với $a - b = n$), tuần $a+1$ đến tuần $b+1$, và tiếp tục, được trình bày bởi một ma trận; các dữ liệu thuê bao sau khi được chuẩn hóa sẽ giúp phân loại các thuê bao thành hai loại: thuê bao giảm tương tác và thuê bao tương tác bình thường; các thuê bao này được đánh nhãn hai lớp kí hiệu là 0 và 1, trong đó 1 là lớp thuê bao giảm tương tác và 0 là lớp thuê bao tương tác bình thường; để xác định một thuê bao là giảm tương tác hay không thì từ dữ liệu chuẩn hóa, dựa trên tổng lượng sử dụng của tuần đó thấp hơn ít nhất 40% tổng lượng sử dụng của tuần liền trước; để xử lý khi có khối lượng rất lớn thuê bao và cần hệ thống chạy theo thời gian thực, có thể phân cụm thuê bao theo tương tác thành một số cụm mà mỗi cụm sự tương tác của thuê bao là tương đồng;

b2: gán nhãn thuê bao;

sau khi chuẩn hóa dữ liệu của các thuê bao, sử dụng công cụ xử lý dữ liệu lớn SparkSQL, Spark để thêm một trường nhãn vào ma trận dữ liệu, trường nhãn này có giá trị 0 hoặc 1 ứng với mỗi dòng dữ liệu đặc trưng cho mỗi thuê bao, trong đó 1 là lớp thuê bao giảm tương tác và 0 là lớp thuê bao tương tác bình thường;

bước 3: huấn luyện mô hình;

dữ liệu sau khi tiền xử lý (chuẩn hóa và gán nhãn) tiếp tục được lưu trữ trên cơ sở dữ liệu của hệ thống và sử dụng trên khối huấn luyện mô hình; các mô hình học máy có thể lựa chọn trong số các mô hình phân lớp như rừng ngẫu nhiên tăng cường độ dốc (XGboost); XGboost là một trong những phương pháp tốt nhất hiện nay trên thế giới được sử dụng rộng rãi trong lĩnh vực phân loại và dự báo từ dữ liệu, vì thế chúng tôi không mô tả chi tiết phương pháp này; nhờ vào phương pháp xử lý dữ liệu và gán nhãn từ bước 2, giúp đưa dữ liệu về dạng chuẩn hóa để có thể dùng cho XGboost; việc huấn luyện mô hình được triển khai trên các máy chủ có khả năng xử lý cao khi dữ liệu thuê bao lên đến hàng triệu thuê bao; bước này giúp tinh chỉnh tham số của mô hình và đầu ra là mô hình tối ưu dựa trên tri thức thu được từ dữ liệu chuẩn hóa, mô hình được lưu lại và được kiểm nghiệm trước khi đưa vào dự đoán thuê bao giảm tương tác;

bước 4: lựa chọn thời điểm T để kiểm nghiệm mô hình và tiến hành kiểm nghiệm mô hình;

để kiểm nghiệm mô hình trên khối kiểm nghiệm, với đặc thù dữ liệu chuỗi thời gian, ta chia dữ liệu dựa trên mốc thời gian tại thời điểm T; thời điểm T được lựa chọn dựa trên tập dữ liệu được thu thập, đảm bảo 75% dữ liệu trước thời điểm T làm dữ liệu huấn luyện và 25% dữ liệu sau thời điểm T làm dữ liệu kiểm nghiệm; hiệu quả sau huấn luyện với cách xử lý dữ liệu như mô tả cho mô hình dự đoán rừng ngẫu nhiên tăng cường, đạt tỉ số F1 xấp xỉ 80%; theo thời gian có thể mô hình sẽ lạc hậu, vì thế hiệu năng sau khi kiểm nghiệm sẽ giảm; lúc này dữ liệu được thu thập thêm và mô hình sẽ được cập nhật và huấn luyện lại, thời điểm T cũng sẽ được cập nhật theo quy tắc 75% - 25%; phần 25% dữ liệu dùng để kiểm nghiệm được đưa vào mô hình và dự đoán nhãn; so sánh nhãn dự đoán bởi mô hình với nhãn thật để tính được độ chính xác của mô hình; độ chính xác cao hơn một ngưỡng cho trước, ví dụ 80% thì mô hình đạt yêu cầu;

sau khi lựa chọn được thời điểm T để kiểm nghiệm mô hình, tiến hành kiểm nghiệm mô hình trên hệ thống máy tính dựa vào kết quả dự đoán của mô hình cho dữ liệu kiểm nghiệm; với mục đích giữ thuê bao người dùng, cần dự báo chính xác người dùng giảm tương tác hơn là dự báo người dùng không giảm tương tác; việc nhầm lẫn một thuê bao tương tác thành giảm tương tác có thể dẫn tới sự kém hiệu quả của các

chương trình tương tác sau này, tuy nhiên bỏ qua một thuê bao giảm tương tác có thể dẫn đến mất đi một người dùng, trực tiếp ảnh hưởng đến hiệu suất; vì vậy độ chuẩn xác của mô hình kiểm nghiệm (F1) sẽ dựa vào các số đo:

$$\text{độ chính xác} = \frac{\text{tổng người dùng giảm tương tác được dự đoán đúng}}{\text{tổng người dùng được dự đoán là giảm tương tác}}$$

$$\text{thu hồi} = \frac{\text{tổng người dùng giảm tương tác được dự đoán đúng}}{\text{tổng người dùng thực tế là giảm tương tác}}$$

và trung bình điều hòa để tính ra độ chuẩn xác của mô hình kiểm nghiệm:

$$F1 = 2 \cdot \frac{\text{độ chính xác} \cdot \text{thu hồi}}{\text{độ chính xác} + \text{thu hồi}}$$

mô hình đạt yêu cầu khi F1 từ 75% trở lên; tùy vào yêu cầu thực tế có thể thay đổi ngưỡng này;

bước 5: dự đoán thuê bao giảm tương tác;

trong bước này, mô hình thu được và được lưu trữ trong bước 3 đã được kiểm nghiệm trong bước 4 được sử dụng tại khối dự đoán; lúc này với dữ liệu thuê bao mới được đưa vào hệ thống ở thời điểm hiện tại, cần dự đoán tuần tiếp theo thuê bao có giảm tương tác hay không; nếu trong bước 3, dữ liệu lịch sử được thu thập để huấn luyện và thu được mô hình, thì bước 5 dùng mô hình từ bước 3 để dự đoán cho dữ liệu mới chưa có trong tập dữ liệu huấn luyện; nếu bước 3 huấn luyện mô hình đòi hỏi hệ thống máy tính mạnh để thực hiện trong một thời gian khá lâu có thể mất vài giờ thì trong bước 5 với hệ thống máy tính tương tự, việc dự đoán sẽ diễn ra nhanh chóng chỉ trong vòng khoảng 15 phút;

hệ thống máy tính sử dụng ở bước 3 và bước 5 gồm 100 máy chủ kết nối với hệ thống xử lý dữ liệu lớn (hadoop), mỗi máy chủ có cấu hình: hệ điều hành Redhat 6.8, RAM; 256 GB, HDD 1.5TB, SSD: 1.5TB, CPU: 48 cores, chip: Intel(R) Xeon CPU ES-2690 v2 3.00 Ghz; chạy trên hệ thống là công cụ Spark trong hệ thống phân tán với tham số 18Gb bộ nhớ thực thi, 10 Gb bộ nhớ quản lý tiến trình, 5 nhân thực thi và 20 máy thực hiện.

